

# Exploiting Online Discussions to Discover Unrecognized Drug Side Effects

H. Wu; H. Fang; S. J. Stanhope

University of Delaware, Newark, Delaware, USA

## Keywords

Algorithms, adverse effects, online discussions

## Summary

**Background:** Drugs can treat human diseases through chemical interactions between the ingredients and intended targets in the human body. However, the ingredients could unexpectedly interact with off-targets, which may cause adverse drug side effects. Notifying patients and physicians of potential drug effects is an important step in improving healthcare quality and delivery.

**Objective:** With the increasing popularity of Web 2.0 applications, more and more patients start discussing drug side effects in many online sources. These online discussions form a valuable source for mining interesting knowledge about side effects. The main goal of this paper is to investigate the feasibility of exploiting these discussions to discover unrecognized drug side effects.

**Methods:** We propose methods that can 1) build a knowledge base for drug side effects by automatically integrating the information related to drug side effects from

different sources; and 2) monitor online discussions about drugs and discover potential unrecognized drug side effects.

**Results:** Experiment results show that the online discussions indeed provide useful information discovering unrecognized drug side effects. We find that the integrated knowledge base contains more information than individual online sources. Moreover, both proposed detection methods can identify the side effects related to the four recently recalled drugs, and the information from online discussions makes it possible to make the detection much earlier than official announcements. Finally, the proposed generative modeling method is shown to be more effective than the discriminative method.

**Conclusions:** We find that it is possible to monitor online discussions to detect unrecognized drug side effects. The developed system is expected to serve as a complementary tool for drug companies and FDA to receive feedbacks from the patients, and it has the potentials to expedite the discovery process of unrecognized drug side effects and to improve the quality of healthcare.

pre-clinical and clinical trial data. Unfortunately, not all side effects can be discovered in a lab test and small clinical trials. Instead, after the drug is approved by U.S. Food and Drug Administration (FDA) and enters the market, all the patients on the market end up being part of a large (post-clinical) experiment to identify unrecognized and unexpected drug side effects. Currently, in such post-market situation, drug side effects are often reported by physicians based on the information gathered from their patients through Med-Watch systems. If the unexpected side effects are dangerous or fatal, FDA or the drug company may decide to withdraw the drug from the market. This process might take up to a few years.

Recently, with the increasing popularity of Web 2.0 technology, more and more patients share their experiences and discussing drug side effects on various online web sites prior to their doctor visits. Such online information forms another valuable information resource about drug side effects. Together with the information from existing systems such as FDA Adverse Event Reporting System (FAERS) [18], these online discussions makes it possible to provide FDA and drug companies with a more complete set of side effect related information and to discover unrecognized drug side effects earlier. Unfortunately, such information is currently under-utilized.

In this article, we study the feasibility of exploiting online discussions to discover unrecognized drug side effects. First, we propose to create a knowledge base for known drug side effects by automatically integrating different online sources. The constructed knowledge base is expected to include information of all drugs and their known side effects, and it enables us to analyze and discover related drug side ef-

## Correspondence to:

Hao Wu  
209 Evans Hall  
University of Delaware  
Newark, DE 19716  
USA  
E-mail: haowu@udel.edu

Methods Inf Med 2013; 52: 152–159

doi: 10.3414/ME12-02-0004

received: March 2, 2012

accepted: November 22, 2012

prepublished: ■■■

## 1. Introduction

Each drug has both benefits and risks. The interaction between a drug and its intended targets can treat the diseases associated with the targets, while the interaction with “off-targets” may make drugs less effective or

even cause dangerous side effects such as heart failure [1–3]. Physicians and patients need to know possible drug side effects in order to reduce serious accidents resulting from adverse drug-related events [4].

In the pre-market situation, some side effects of a drug can be recognized in the

fects. Second, we propose two methods that can detect potential unrecognized drug side effects based on online discussions about drugs. The first one is a discriminative classification method based on accumulated discussion frequency to decide whether a side effect is likely to be associated with a drug. The second one is a generative modeling method that directly estimates the probability a drug may be related to a side effect based on the discussions.

We conduct three sets of experiments to validate the proposed integration and detection methods. Our study suggests that online discussions can provide valuable information about drug side effects and should be utilized and combined with the information collected by existing systems to develop a more effective early warning system for FDA and drug companies to detect unrecognized adverse side effects.

## 2. Related Work

Web 2.0 technologies enable online users to share their opinions and experiences about health-related issues. As a result, social media has been shown to be a useful source for mining health-related information. For example, Heavilin et al. conducted a study on dental-related tweets and found that people extensively shared their experiences related to their dental visits and pains [15]. Chew et al. analyzed the tweets during the 2009 H1N1 outbreak, and found that the frequency of H1N1 related tweets was highly correlated with the number of H1N1 news events [13]. Signorini et al. used the tweets during H1N1 outbreak to estimate the disease level [17]. Paul et al. demonstrated that discussions from tweets were useful to locate geographic syndrome [16].

There have been a few studies that utilized online information for side effect related problems. Carlo et al. studied the problem of mining unrecognized side effects from the Web pages [5]. They proposed to use neural networks to decide whether a web page mentions any drug side effects, and used FDA labels and MedlinePlus as references for known side effects. Our work focuses on a different

source of information, i.e., online discussions. Compared with Web pages, online discussions are user-generated content, which means the quality of the information might be lower but the discussion volume would be much higher. Leaman et al. demonstrated the feasibility of extracting adverse drug side effects from user comments [11]. However, they did not focus on finding unrecognized drug side effects. In particular, their evaluation is based on whether the most frequent side effects found from the user comments match the most common known side effects of the drug. Moreover, they only focused on a single online source, i.e., DailyStrength. On the contrary, we aim to fully exploit online discussions from multiple sources. More recently, Chee et al. studied how to predict possible candidate recalled drugs based on online health forum [10]. The problem is different from what we study in this article, i.e., whether a drug might be related to a side effect.

This article is in the similar direction as the FDA Sentinel Initiative, which aims to “query diverse automated healthcare data holders ... to evaluate possible medical product safety issues quickly and securely” [19]. The similarity is that we also try to implement a system that complements existing systems on tracking adverse event reports. The difference is that the Sentinel systems focuses on official records such as electrical health record systems, administrative and insurance claims databases, while we focus on online discussions of the patients.

## 3. Our Methods

### 3.1 Constructing a Drug Side Effect Knowledge Base

Since our goal is to discover unrecognized side effects of drugs, we first need to build a knowledge base that includes most of the known drug side effects. Specifically, the knowledge base should include information about as many drugs and their associated side effects as possible.

A few online sources provide information related to drug side effects. For example, *SIDER* contains extracted drug side effects from public documents and

provides the information in a well-structured format [9]. *DailyMed* provides high quality information about drugs approved by FDA including FDA labels. *Drugs.com* is one of the most visited drug-related web sites. By comparing the information from these three sources, we find that none of them contain all the drug-related information. Moreover, we notice that the languages used to describe side effects are different in different sources. For example, the terminologies used in *DailyMed* are often more formal since it comes from the drug label, while the languages used in *Drugs.com* are more conversational since they come from the patients. Thus, it is necessary to integrate the information from all these sources to construct a more complete knowledge base.

Among all the three sources we consider in the article, only *SIDER* provides structured information that makes it possible to directly extract drug names and side effects. Unfortunately, the other two sources are unstructured, so it is more challenging to extract drug names and side effects from them. We notice that most pages from these three sources are organized based on drugs. Every page discusses the information of a single drug, and drug names are often mentioned in specific fields such as “title”, “drug” or “drug name” in the HTML pages. Thus, a simple yet effective drug name extraction strategy is to utilize the HTML template of each web source, identify the field related to drug names, and use these field values as drug names.

Unlike drug names that are often the values of specific fields, side effect names are scattered in the plain text with noisy terms such as drug descriptions or drug labels. Thus, the drug name extraction method described above would not work well for side effect name extraction. To solve the problem, we propose to use a lexicon to extract drug side effect names from these plain text. In particular, we use the side effect names from *SIDER* as the lexicon. *SIDER* is one of the most representative resources about drug side effects [9], and it contains 1,450 side effect names. Note that we can easily include new side effects that are not from *SIDER* by adding them to the lexicon. It is clear that the effectiveness of our matching method is not

$$P(D(d,e) = 1|D(d) = 1) =$$

$$P(R(d,e) = 1|D(d) = 1) \times P(D(d,e) = 1|R(d,e) = 1, D(d) = 1) \\ + P(R(d,e) = 0|D(d) = 1) \times P(D(d,e) = 1|R(d,e) = 0, D(d) = 1)$$

**Figure 1** The probability of observing a discussion about a drug (d) and a side effect (e)

limited by the completeness of the SIDER database.

Given the lexicon (i.e., a list of drug side effect names), we then use them to match the pages from those online sources and decide whether a page is about the drug side effect. Instead of using only exact matching for side effect names, we pre-process the documents using Porter Stemmer, which normalizes the terms and makes it possible to match terms with the same stem form, e.g., “fevers” and “fever”. Moreover, instead of using exact string matching, we compute the similarities between strings based on their overlapped terms. This strategy would allow us to identify variants of a side effect such as “lung cancer” and “cancer of lung”.

After extracting drugs and side effects from the three different sources, we can then construct an integrated knowledge base of drug side effects with a list of drugs and their associated side effects. The comparison of the knowledge base with individual sources is discussed in Section 4.

## 3.2 Detecting Unrecognized Drug Side Effects

Given a drug and a side effect, the problem is to predict whether the side effect is indeed related to the drug based on online discussions. It is clearly a binary classification problem. If the side effect is related to the drug, we classify the drug side effect as “relevant” to the drug. Otherwise, the drug side effect will be classified as “non-relevant”. To solve the problem, we propose a discriminative classification method and a generative modeling method.

### 3.2.1 Discriminative Classification Method

Intuitively, if the side effect is indeed associated with the drug, more people will mention it in the online discussions. Thus, “relevant” side effects should have higher discussion frequency than the “non-relevant” side effects. A commonly used classification method is discriminative methods with the goal of directly modeling the boundary between the two categories. In this article, we use Rocchio method [6], where the basic idea is that the label of a new data point is decided based on the distance of the data point to the centroid of each category.

Specifically, given a drug, we construct a training data set based on the information about the drug from the knowledge base. For all of its known side effects, we first collect online discussions for each of them and then compute their average discussion frequency. We repeat the same procedure for the unknown side effects of the drug (i.e., those are not associated with the drug based on the information from the knowledge base). Now we want to consider whether a side effect is “relevant” to the drug. To answer the question, we first compute its discussion frequency and then compare it with the average frequency of known side effects and that of unknown side effects. If it is closer to the average discussion frequency of the known side effects, this side effect will be classified as “relevant”, i.e., the ones associated with the drug. Otherwise, the side effect will be classified as “non-relevant”. For all the side effects classified as “relevant”, we will check the knowledge base to see whether it has been recognized. If not, we will label this side effect as a potential unrecognized one.

### 3.2.2 Generative Modeling Method

One limitation of the discriminative method is that it assumes the discussion frequency is the only factor deciding whether a side effect is associated with a drug. If there are more discussions about both a drug and a side effect, it is more likely to conclude that the side effect is indeed related to the drug. However, this assumption does not hold in reality. For example, common side effects, such as “fevers”, are more likely to be discussed than the rare ones.

In fact, there could be two possible scenarios when a user mentions a side effect when he or she discusses a drug: 1) the user took the drug, suffered the side effect, and wanted to share his experience; or 2) the user did not take the drug or suffer the side effect, and simply mentioned the side effect for other reasons, such as consulting before taking the drug. It is important to distinguish these two scenarios since the latter one clearly does not provide reliable information about the relatedness of the side effect and the drug. Thus, we propose a generative model that can better describe the generation process of the online discussions.

Let us first explain the notations.

- d: a drug;
- e: a side effect that might be related to d;
- $D(d)$ : a binary random variable indicating whether there is a discussion about drug d;
- $D(d,e)$ : a binary random variable indicating whether there is a discussion about drug d and side effect e;
- $R(d,e)$ : a binary random variable denoting whether e is indeed related to d or not.

We now describe the basic idea. Given that a user discusses a drug (i.e.,  $D(d) = 1$ ), we now describe how a discussion about the drug and a side effect is generated. The first possible generation process is that the user experienced the side effect and then mentioned it in the discussion, and the second is that the user did not experience it and mentioned it for other reasons. Thus, we can write down the probability of observing a discussion about d and e as can be seen in ► Figure 1.

**Figure 2**

The likelihood of the side effect is related to the drug given the discussions

$$P(R(d, e) = 1 | D(d) = 1) = \frac{P(D(d, e) = 1 | D(d) = 1) - P(D(d, e) = 1 | R(d, e) = 0, D(d) = 1)}{P(D(d, e) = 1 | R(d, e) = 1, D(d) = 1) - P(D(d, e) = 1 | R(d, e) = 0, D(d) = 1)}$$

Our goal is to estimate how likely the side effect is related to the drug given the discussions, so we need to compute  $P(R(d, e) = 1 | D(d) = 1)$ . Since we have  $P(R(d, e) = 1 | D(d) = 1) + P(R(d, e) = 0 | D(d) = 1) = 1$ , we can re-writing the above question and get the result shown in ► Figure 2.

We now discuss how to estimate each component in the above equation. The first component is  $P(D(d, e) = 1 | D(d) = 1)$ , and it can be directly estimated as follows:

$$P(D(d, e) = 1 | D(d) = 1) = \frac{\text{\# of discussions of drug } d \text{ mentioning side effect } e}{\text{\# of discussions about drug } d}$$

The second component is  $P(D(d, e) = 1 | R(d, e) = 1, D(d) = 1)$ , i.e., the probability that a user would mention the side effect when the user is discussing the drug and has experienced the side effect caused by the drug. Intuitively, the probability would be very high since people like to share their experience. In this article, we set the probability to 1. We also tried other values, and found that the value does not affect the performance significantly.

The last component is  $P(D(d, e) = 1 | R(d, e) = 0, D(d) = 1)$ , i.e., the probability that a user would mention the side effect randomly in the discussion of the drug. If we have data specifically from this scenario, we could directly compute the probability based on the training data. However, we do not have the data because it is impossible to reach out to every online user and ask about their intentions when writing the discussions. Therefore, without loss of generality, we assume that this probability is drug-independent, which means that we have the probability of a side effect being mentioned randomly in a discussion about a drug is the same for all drugs. Thus, we have

$$P(D(d, e) = 1 | R(d, e) = 0, D(d) = 1) = \frac{\text{\# of discussions mentioning side effect } e}{\text{\# of all the discussions in the collection}}$$

## 4. Results

### 4.1 Effectiveness of the Proposed Integration Method

► Table 1 shows the statistics of the three sources and our integrated knowledge base. It is clear that our knowledge base contains a more complete set of information about drug side effects than each of the individual sources. In particular, the integrated knowledge base contains the information about more drugs and more side effects.

We also show the number of overlapped concepts (i.e., drug and side effect pairs) for each pair of resources in ► Table 2. Moreover, ► Table 3 shows example side effects of drug “deferasirox” from different

sources. Note that a side effect could occur in more than one source, and the table only shows unique side effects from each single source. It is clear that different sources contain different side effects or different ways of describing side effects, and it is necessary to integrate the information from multiple sources.

### 4.2 Effectiveness of the Proposed Detection Methods

To detect unknown side effects from online discussions, we first need to crawl online discussions about drug related information. One possible strategy is to start with a list of health-related blogs and crawl information from these web sites. However, the coverage of these blogs is often limited. Thus, we explore another solution in the developed system. In particular, we rely on forum search engines such as Google Discussion Search and crawl all drug related information. In particular, we formulate

**Table 1** Statistics of the three sources and the integrated KB

	SIDER	Drugs.com	DailyMed	Our KB
# of drugs	4,953	11,652	12,865	15,848
# of drug/side-effect pairs	323,198	555,978	83,848	842,955

**Table 2**

The number of overlapped concepts (i.e., drug/side-effect pairs)

	SIDER	Drugs.com	DailyMed
SIDER	323,198	74,081	38,656
Drugs.com	74,081	555,978	25,464
DailyMed	38,656	25,464	83,848

**Table 3** Example side effects of drug “deferasirox” from different sources

	SIDER	Drugs.com	DailyMed	Our KB
# of side effects	72	63	47	103
Example side effects	Ear pain, abdominal upper pain, ...	Black stools, bleeding, bruising, ...	Duodenal ulcer, erythema, gallstones, ...	All of them from the four sources



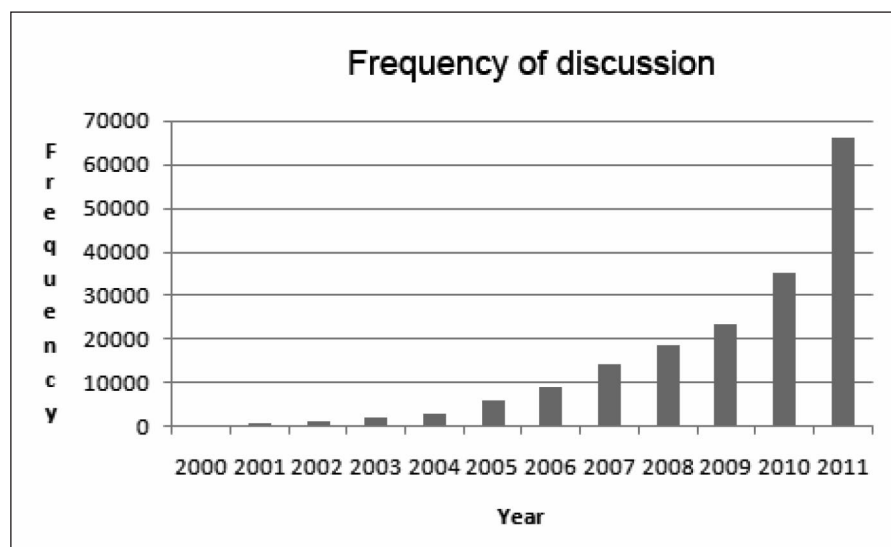


Figure 3 Temporal distribution of our data collection

queries based on both drug and side effect names. The returned results for all the queries are pooled together, and form the online discussion collection used in the experiments. We crawled 178,871 online discussions, and the temporal distribution is shown in ►Figure 3. The figure shows a clear trend that the number of online discussions about drug side effects has exponentially increased in recent years. Before 2005, there were less than 5,000 discussions

per year. In 2011, the number becomes more than 65,000.

We conduct two sets of experiments to evaluate the effectiveness of the proposed detection methods. First, we evaluate the effectiveness through retrospective evaluation [8]. In particular, we picked four drugs that have been recently recalled because of some unrecognized side effects, and see whether the proposed methods are able to identify these side effects. Second,

we apply the proposed methods to monitor the online discussions about twenty popular drugs, and compare the effectiveness of the proposed two methods in predicting possible unrecognized side effects that have been reported in the FAERS system.

#### 4.2.1 Retrospective Evaluation

One way of evaluating the effectiveness of the proposed methods is to see whether they can identify unrecognized drug side effects. To achieve this goal, we select four drugs that have been recently warned or recalled by FDA, and see whether the proposed methods can identify those side effects causing the recall or warnings. The four drugs are Darvocet, Yaz, Yasmin and Zocor. We crawl the online discussions about these drugs and apply the proposed detection methods.

Let us take Darvocet as an example. It was recalled by FDA on Nov. 19, 2010 for its risk of abnormal heart rhythms which may cause sudden death. We collected the online discussions about the side effect. ►Figure 4 shows the number of discussions about the side effect over recent year. It is clear that there have been quite a few discussions about the side effect since 2006, four years earlier than the official recall. We then applied our proposed method, and both of them can effectively predict that the side effect is associated with the drug much earlier than the official announcement.

Specially, ►Figure 4 illustrates the results of the discriminative method. The x-axis is the timeline, and the y-axis is the discussion frequency. The lines labeled as “known” and “unknown” represent the average accumulated discussion frequencies for known and unknown drug side effects, respectively. The line “threshold” is in the middle of these two lines, and indicates the classification boundary. Given a time stamp, if the accumulated discussion frequency of the side effect is larger than the corresponding value at the classification boundary, the side effect will be predicted as “relevant” to the drug. ►Figure 4 shows that the discussion frequency line of side effect “heart rhythm” is way above the classification boundary (i.e., threshold line), and this trend has become obvious since

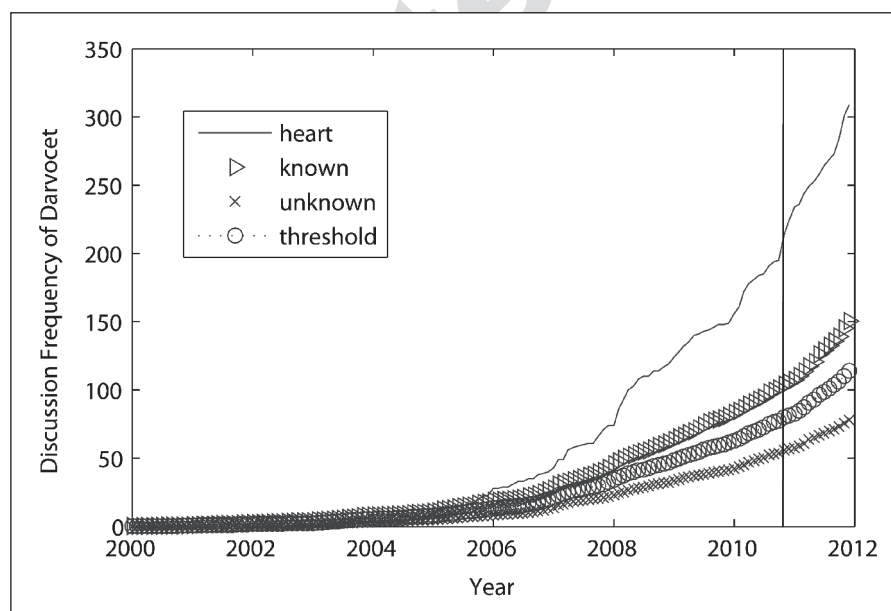
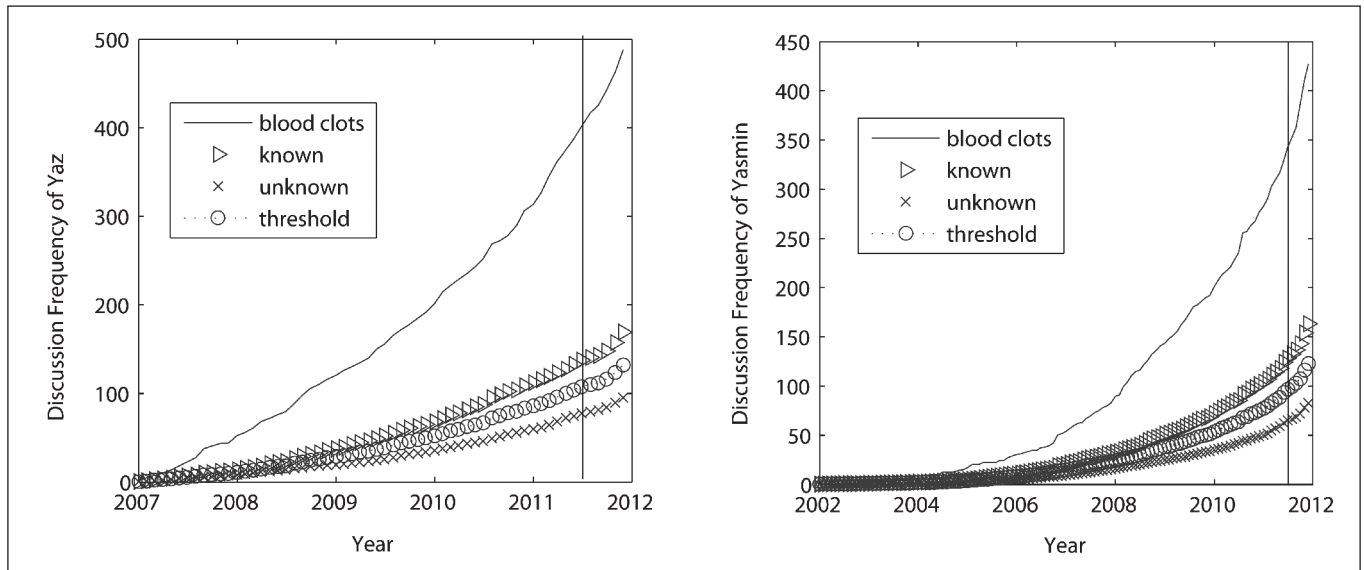


Figure 4 The accumulated discussion frequency and results of the discriminative method for Darvocet (Darvocet: abnormal heart rhythms. Recalled by FDA on Nov. 19, 2010)



**Figure 5** The accumulated discussion frequency and results of the discriminative method (Yaz (left) and Yasmin (right): blot clots. Warned by FDA on May 31, 2011)

2006. The detailed information about the frequency and decision boundary is shown in ► Table 4.

We conduct similar experiments for the other three drugs, i.e., Zocor (FDA issued warning for the side effect of muscle damage on 03/19/2010), Yaz and Yasmin (FDA issued warning for the risk of blood clots on 05/31/2011 for both drugs). Our results show that both discriminative and generate methods are effective to detect these drug side effects 4–6 years earlier than the official warnings. The results for these drugs are shown as ► Figure 5.

#### 4.2.2 Comparison with FAERS Reports

Although the results in the retrospective mining experiments show that the proposed methods are effective, it still remains unclear which one of the proposed methods is more effective.

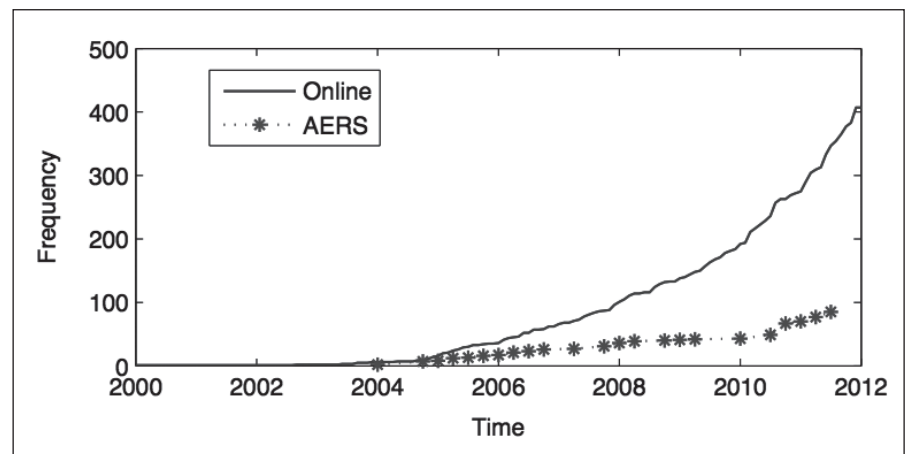
To quantitatively compare the two methods, we conduct another set of experiments by leveraging FAERS, i.e., a database with drug side effect related reports that have been submitted to FDA. It contains the information about drug side effects gathered from a different channel than ours, and it would be interesting to leverage the database to compare our methods. Note that the information from the FAERS data-

base is different from the FDA official drug recall/warning information as follows. FAERS maintains a record of side effect cases, which are utilized by FDA to make the official recall/warning decisions. This information is reported by physicians or patients, but the side effect is not confirmed until the official announcements by drug companies or FDA.

We conduct a set of experiments to examine whether the side effects discovered based on online discussions are consistent with those reported in FAERS. Thus, we need to compare the drug side effects detected by our methods with those re-

ported in the FAERS system. In particular, we conduct experiments over 20 popular drugs selected based on drugs.com, and the drugs are as following: *Ambien, Amitriptyline, Amoxicillin, Atenolol, Ativan, Cephalixin, Clonidine, Flexeril, Gabapentin, Hydrochlorothiazide, Klonopin, Lexapro, Methocarbamol, Metoprolol, Morphine, Neurontin, Phentermine, Seroquel, Skelaxin, Soma*.

For each drug and an unknown side effect pair, if the number of FAERS reports is more than 10, we assume that the side effect is “relevant” to the drug according to FAERS. Based on the assumption, we may



**Figure 6** Comparison of Discussion Frequency from online discussion vs. FAERS reports

**Table 4** Accumulated discussion frequency and the centroids of positive/negative examples for Darvocet. (positive examples: known side effects; negative examples: unknown side effects)

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011
Discussion Frequency of "abnormal heart rhythm"	3	7	23	40	74	119	149	224	309
Decision Boundary	5.5	7.5	13	20	31.5	47	61	81	114
Center of Positive Examples	8	10	18	27	43	62	81	105	150
Center of Negative Examples	3	5	8	13	22	32	41	57	78

construct a list of FAERS-relevant side effect for each drug, and these lists are used as judgments.

The evaluation measure we used here is precision and recall, which are basic measures used in Information Retrieval. In particular, precision is to measure the percentage of predicted drug side effects that are covered by FAERS. And it can be computed through dividing the number of drug side effects that are both discovered by our methods and reported in FAERS system with the number of drug side effects discovered by our methods. On the other hand, recall is to measure the percentage of drug side effects reported in FAERS that are also predicted by the method. And it is computed through dividing the number of side effects that are both discovered by our method and reported in FAERS system with the number of side effects from the FAERS system.

► Table 5 shows that the evaluation results of the proposed to methods based on FAERS judgments. It is clear that the generative model method is more effective than the discriminative classification method in terms of both precision and recall. This is expected since the assumption made in discriminative method is too strong, and the generative model is designed to break this limitation. Moreover, the results indicate that online discussions are a good source for drug side effect related information, and a majority of discovered side effects are consistent with those reported in FAERS system.

We believe that the online discussions are a good complementary source for existing FAERS system because the discussion volume is often large and the feedback time is often shorter than the FAERS reports. To verify our hypothesis, we compare the information from online discussions and

FAERS system for each pair of drug and unrecognized side effect. ► Figure 6 shows an example plot about the side effect "muscle injury/weakness" of the drug "Zocor". It is clear that the number of reports from online discussions is much larger than those from FAERS system.

### 4.3 Discussion

One of our assumptions is that all discussions about a drug and a side effect can be used to confirm their association. However, the assumption would not always hold since the discussions may convey negative meaning. For example, a user may mention that he or she does not have a side effect. If such cases happen frequently in our data set, the results of the proposed methods would not be valid since a discussion about not having a side effect might be mistakenly considered as the one mentioning the side effect.

To examine how often our assumption holds in the real-world data, we manually checked the discussions of a few drugs and found that the number of discussions with negative meanings (i.e., the user does not have the side effect) is rather small and does not affect the prediction results. For example, let us consider the drug "Darvocet". There are 6,344 discussions about the drug in our collection, and we find 1,344 discussions with negative terms such as "never", "not", "no", "cannot", etc. We then

manually go through each of the discussion, and found that only nine discussions with real negative meanings (i.e., the user does not have the side effect). Here are a couple of examples with no real negative meanings: "It's not fun to vomit when your mouth hurts so bad..." and "it can trigger depression, not in everyone of course, but I ...". Among the nine discussions with real negative meanings, six of them are about known drug side effects (i.e., constipation, nausea and swelling), and only three of them are about unknown drug side effects (i.e., cough and fever).

In summary, we conduct manual experiments to examine how the discussions with negative terms would affect the performance, and we can show that the percentage of discussions with real negative meaning is rather small and does not affect the prediction performance.

## 5. Conclusions and Future Work

Patient's discussions on the World Wide Web are a valuable source to discover recognized drug side effects. To take advantage of the source in discovering unrecognized drug side effects, we propose two methods: discriminative classification and generative modeling. Experiment results show that 1) online discussions are useful to detect unknown drug side effects; and 2) the generative modeling method is more effective in terms of both precision and recall. We believe that our work is a good complementary of existing drug side effect discovery system and could be leveraged as additional evidence in the discovery process.

In the future, we plan to continue our work in the following directions. First, we

**Table 5** Comparison of the two methods based on the coverage of FAERS reports

Methods	Precision	Recall
Discriminative Classification	0.69	0.23
Generative Model	0.70	0.69

plan to explore more advanced natural language processing methods to decide whether a discussion indeed talks about the side effect of the drug. Second, we will study how to systematically combine information from different sources such as on-line discussions, scientific literature, news articles, FAERS reports to make more accurate prediction.

## Acknowledgments

This material is based upon work supported by the University of Delaware Research Foundation Strategic Initiatives and the National Science Foundation under Grant Number IIS-1017026. We thank the reviewers for their useful comments.

## References

- Bertsche T, Pfaff J, Schiller P, Kaltschmidt J, Pruszydio MG, Stremmel W, Walter-Sack I, Haefeli WE, Encke J. Prevention of adverse drug reactions in intensive care patients by personal intervention based on an electronic clinical decision support systems. *Intensive Care Medicine* 2010; 36 (4): 665–672.
- Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug Target Identification Using Side-Effect Similarity. *Science* 2008; 321: 263–266.
- Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. *The Lancet* 2000; 356 (9237): 1255–1259.
- Berry DC, Knapp P, Raynor DK. Provision of information about drug side-effects to patients. *The Lancet* 2002; 359 (9309): 853–854.
- Curino CA, Jia YY, Lambert B, West PM, Yu C. Mining officially unrecognized side effects of drugs by combining web search and machine learning. In *Proceedings of the 14th ACM international conference on Information and knowledge management* 2005; 365–372.
- Rocchio, JJ. Relevance feedback in information retrieval. In: Gerard S, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*. New Jersey: Prentice-Hall; 1971. pp 313–323.
- Cover T, Hart PE. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 1967; 13 (1): 21–27.
- Yang Y, Pierce T, Carbonell JG. A study on retrospective and on-line event detection. In: *Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval* 1998. pp 28–36.
- Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork PA. Side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology* 2010; 6: 343.
- Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. In: *Proceedings of AMIA Annual Symposium* 2011. pp 217–226.
- Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In: *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing* 2010. pp 117–125.
- Kamel Boulos MN, Sanfilippo AP, Corley CD, Wheeler S. Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance. *Computer Methods and Programs in Biomedicine* 2010; 100 (1): 16–23.
- Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PLoS ONE* 2010; 5 (11).
- Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS ONE* 2011; 6 (8).
- Heaivilin N, Gerbert B, Page JE, Gibbs JL. Public health surveillance of dental pain via Twitter. *Journal of Dental Research* 2011; 90 (9): 1047–1051.
- Paul MJ, Dredze M. You are what you tweet: analyzing twitter for public health. In: *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media* 2001. pp 265–272.
- Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS ONE* 2011; 6 (5).
- FDA Adverse Event Reporting System (FAERS). <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm>
- FDA's Sentinel Initiative. <http://www.fda.gov/Safety/FDASentinelInitiative/default.htm>